DataDEX.trade 基于联合计算的去中心化数据交易所(draft-v0.2)

1 背景

传统的互联网模式，用户隐私保护倍受挑战。在中心化的平台和业务上，个人隐私的数据如何被使用是不可监管的，更不公平，数据的所有者没有作为主体，数据被平台方以各种途径使用。而且由于是中心化的数据管理，数据交换的门槛会很高，话语权在个别中心化的巨头手里，流动性差，缺乏统一安全标准。

另一方面，目前的智能设备分布越来越广泛，这些越来越强大的计算资源分散和孤立存在，并且这些算力是低功耗、高性能的，只需要在日常环境下即可工作。加之目前共有云大数据处理成本相对昂贵，不能满足社会化使用数据的需求。

在数据被滥用和算力闲置普遍的背景下，去中心化的共享计算和数据交换成为解决市场痛点的方案。通过建立统一的低成本计算平台，建立共同参与的共享机制，在区块链的全民“持股”的激励模式推动下，形成数据可信交换和安全使用的价值平台。

1.1 数据的价值

在互联网的环境下，用户的数据在使用平台的过程中，个人信息和行为数据被收集在平台端。这些数据在后端被怎么处理了呢？是否是真的 “取之于用户，用之于用户” ？

- 数据被用于程序化广告，使得广告投放更精准，广告主的转化率提高
- 数据被用于精准推荐内容，使得平台方通过准确的撮合获得更多的服务费用。
- 数据被用于程序化广告后产生交易，平台方和品牌商都受益，而且因因此商品产生溢价，用户承担更多的费用
- 数据被用于更广泛的外部合作，使用户在其他渠道同样被程序化营销
- 数据被他方利用建立新的业务平台

宗上述，用户的数据被利用，并且花时间看广告，拿钱消费，不但核心资产被无偿使用，而且没有任何回报，只是得到了软件的使用权。

1.2 移动端计算能力

ARM 架构的 RISC 芯片计算能力发展迅速，以骁龙为例，平均每两年芯片性能翻一倍。以下是骁龙 845 对比 835 时隔一年的产品迭代后的性能提升。

Geekbench 4 - Floating Point Performance Single Threaded			
	Snapdragon 845	Snapdragon 835	% Increase
SGEMM	16.6 GFLOPS	11.4 GFLOPS	**45.1%**
SFFT	4.23 GFLOPS	2.86 GFLOPS	**47.9%**
N-Body Physics	1400 Kpairs/s	872.2 Kpairs/s	**60.5%**
Rigid Body Physics	8524.2 FPS	6130.5 FPS	**39.0%**
Ray Tracing	354.0 Kpixels/s	232.7 Kpixels/s	**52.1%**
HDR	11.9 Mpixels/s	8.31 Mpixels/s	**43.2%**
Gaussian Blur	34.5 Mpixels/s	23.9 Mpixels/s	**44.3%**
Speech Recognition	17.9 Words/s	13.6 Words/s	**31.6%**
Face Detection	752.4 Ksubs/s	532.8 Ksubs/s	**41.2%**

图 1CPU 性能提升

GPU 的发展速度更快，目前 NVIDIA 最快显卡的 1/10 算力，并且功耗持续降低，价格更为低廉。

GFXBench Manhattan 3.1 Offscreen Power Efficiency (System Active Power)				
	Mfc. Process	FPS	Avg. Power (W)	Perf/W Efficiency
Qualcomm QRD (Snapdragon 845)	**10LPP**	**60.90**	**~4.38**	**13.90 fps/W**
Galaxy S8 (Snapdragon 835)	10LPE	38.90	3.79	10.26 fps/W
LeEco Le Pro3 (Snapdragon 821)	14LPP	33.04	4.18	7.90 fps/W
Galaxy S7 (Snapdragon 820)	14LPP	30.98	3.98	7.78 fps/W
Huawei Mate 10 (Kirin 970)	10FF	37.66	6.33	5.94 fps/W
Galaxy S8 (Exynos 8895)	10LPE	42.49	7.35	5.78 fps/W
Galaxy S7 (Exynos 8890)	14LPP	29.41	5.95	4.94 fps/W
Meizu PRO 5 (Exynos 7420)	14LPE	14.45	3.47	4.16 fps/W
Nexus 6P (Snapdragon 810 v2.1)	20Soc	21.94	5.44	4.03 fps/W
Huawei Mate 8 (Kirin 950)	16FF+	10.37	2.75	3.77 fps/W
Huawei Mate 9 (Kirin 960)	16FFC	32.49	8.63	3.77 fps/W
Huawei P9 (Kirin 955)	16FF+	10.59	2.98	3.55 fps/W

图 2GPU 性能提升和能耗下降

同时手机中的存储能力、基于 h265 的视频解压能力，使得手机的综合算力越来越强大。

1.3 数据交换的障碍

1.3.1 数据不能上链

- ✧ 存储成本问题
- ✧ 数据规格不一致

1.3.2 数据所有权问题

数据不能确权追溯，无法以资产化的形式在平台上流动。由于可复制性，不能保证数据被合法利用。另外第三方平台受业务模式影响，不会把数据方便的返还给客户。

1.3.3 价值问题

如何确定数据的价值是个难题，数据不能以量来确定价值，而是数据在业务中起到了什么作用，并且通过规模化应用建立市场机制，由供需来确定价值。

2 市场情况

2.1 云计算市场规模

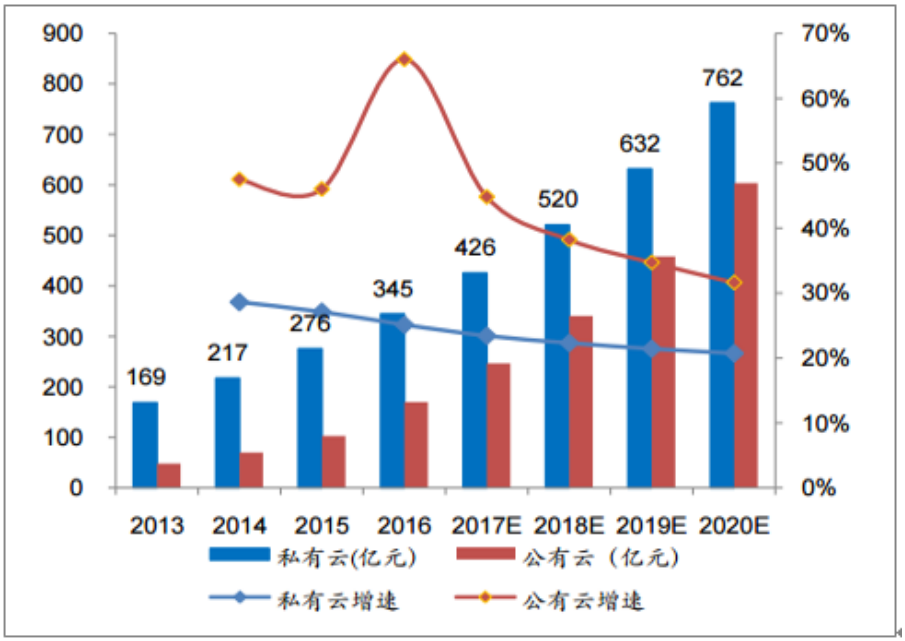


图 3 市场规模预期

从 2017 年开始，公有云市场规模持续增长，但增速下降，并且与私有云的差距缩小。意味着市场对公有云进入理性认知阶段，公有云的市场认可度提高，新入云的客户对公有云的顾虑在减弱。随着安全性、可靠性和信赖程度提高，公有云产品逐渐成为主流。

2.2 大数据市场发展

据中商产业研究院发布的《2018-2023 年中国大数据行业发展前景及投资机会研究报告》数据显示，2017 年中国大数据行业市场规模为 3615 亿元。随着一系列政策的出台，大数据国家战略正在加速落地，大数据行业将持续增长，预计 2018 年中国大数据行业市场规模将近 6000 亿元。

在数据量和数据需求日益增加的情况下，企业的大数据赋能需求日益增加，但是价格逐渐成为中小微企业的门槛。


存储费用

0 ~ 100 GB	大于100GB部分	大于1TB部分	大于10TB部分	大于100TB部分	1PB以上部分
0.0192元/GB/天	0.0096元/GB/天	0.0084元/GB/天	0.0072元/GB/天	0.006元/GB/天	请通过工单联系我们

存储相关说明：收费方式与按I/O计费相同

计算费用

资源定义	CPU	内存	价格
1 CU	1 核	4 GB	150 元/月

图 4 阿里云大数据计算收费

以每年 100T 数据量和 50 个 CPU 计算，每年支出超过 30 万。对于小微企业是个很高的门槛。

2.3 数据交易行业情况

- 众包模式，把数据任务分发，收集后出售
- 授权合作模式，由平台方和所有者转让数据使用权，根据应用效果分成或按量付费
- 公开交易，目前有些灰色，有价值的数据在按潜在规则流转。一些公共数据可利用程度却不高。模式很难大规模运转，用户参与度有限。

- 数据成品交易，以加工后的数据标准格式流转
- 数据应用和报告，以报告或应用产品形式提供服务，形式比较固定，不能灵活应用到其他领域。

目前看来，各类数据交换模式都没有形成社会化规模，缺乏统一的安全控制和激励机制。

3 数据方案

3.1 个人数据确权

大规模的个人数据已成为互联网新的发展动力[1]。但是，个人数据目前被巨头垄断，这不仅阻碍了创新，而且还增加了个人隐私泄露风险，以及隐私数据的滥用。虽然有法律规定保护，但还没有很好的技术基础设施是的监管部门轻松的履行监管。

因此我们提供用户归类和确权个人隐私数据，并且做到细粒度行列级别的确权和访问控制。我们只对数据归属进行确权，而不收集明细数据，因为我们基于联合计算[1]网络，使得以后使用这些已经确权的数据，在用户的主权设备上即可。

3.1.1 目的

我们为了建立了一种可信任使用的大规模个人数据的基础设施，该系统提供了新的个人数据使用范式，将个人数据的所有权和使用权分离，所有权需要通过注册确权，而使用权会在去中心化交易所交易变现。所以我们开发了数据类目确权工具，用户通过安装一个客户端，将个人数据按照统一类目编排，可以选择注册登记以达到确权目的。

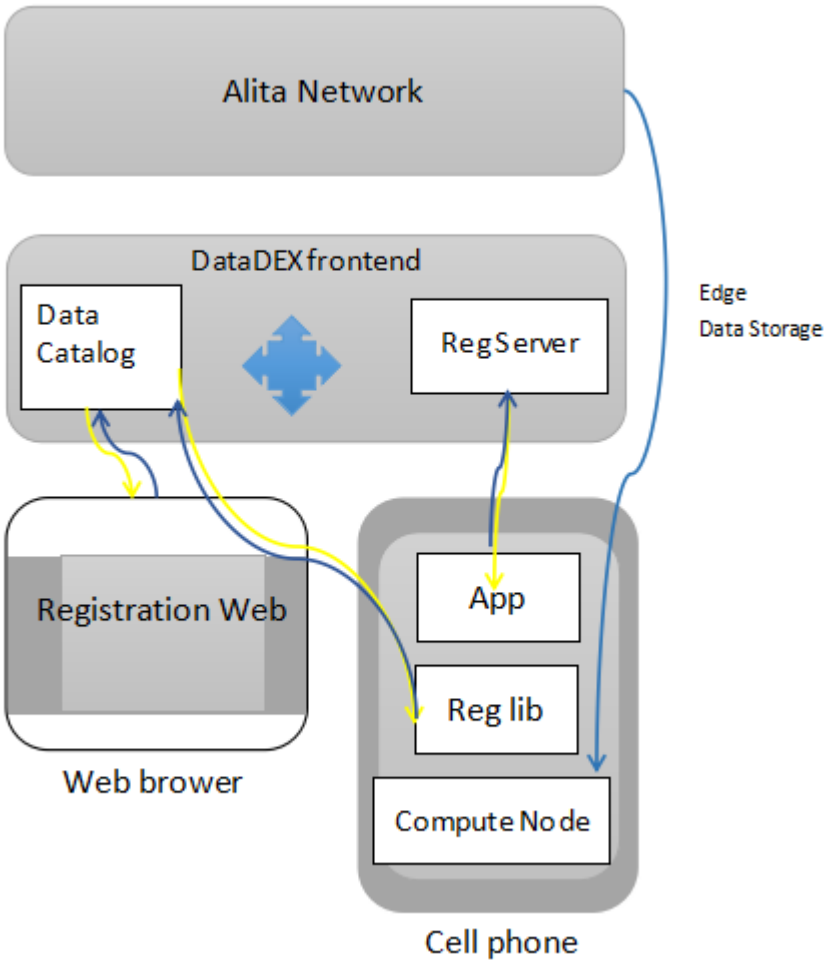
3.1.2 个人隐私数据确权

实施这些数据确权是通过一系列软件实现。用户可以管理软件检测出来的结构化数据，并选择登记确权。用户每天的输出产出条目很多，因此这些软件成为管理这些数据以及登记状态的中心。在软件中允许用户轻松控制数据类目和注册，并管理针对在交易数据集的细粒度授权，以实现新的个人数据价值化愿景。根据公平信息原则，基于已确权的数据交易非常有效，因为用户可以控制对他的数据的访问的人，通过 Data Token。用户可以决定这些数据加入哪些数据集，电商数据是否加入对应的 Dataset，通讯录是否加入对应的 Dataset。

因为用户的编目和数据使用者的编目一致，所以基于联合计算网络，可以无缝地

将新服务与他的数据进行对接，并且不会失去其个人数据的所有权或控制权，具有隐私保护能力。

由于数据确权和交易都能通过软件控制，所以整体流程非常高效和便利。确权系统消除了新企业的准入门槛，它允许更多创新公司提供更好的数据驱动服务。用户选择的服务将有权访问历史数据，甚至在创建服务之前就已经通过持有 Data Token 获得了数据的使用权。而且，明细数据不会被收集，同时也可以通过持有其他 Data Token 获得更丰富的数据。因此，服务提供商可以专注于设计算法向用户提供最佳体验。例如，电商或信息流的个性化推荐。



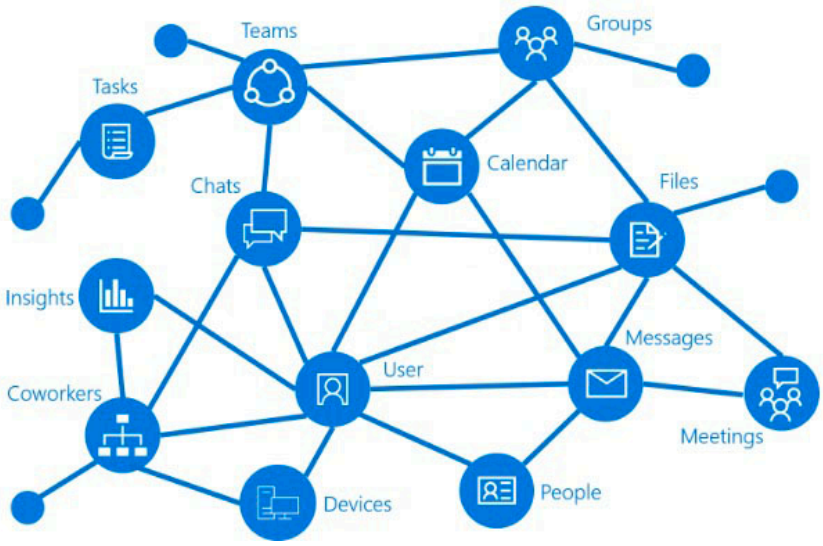
3.1.3 结果

在现有的移动设备中，个人数据已从移动设备中心化服务器上供应用程序使用。这种模型使用户无法控制自己的数据。一旦他们的数据被收集，很难或不可能反驳或撤回。

DataDEX 模型的关键创新在于，可以安全地对用户数据进行计算，在用户控制下的联合计算环境。数据访问权被交易并价值最大化，数据在边缘端计算。同时
用户可以随时查看数据的访问记录，以及数据 Data token 支付情况。

3.2 数据类目

我们第一个阶段针对个人隐私数据，并不是商业数据，所以不像 Alibaba 的 One Data[3]采用数据仓库的模型设计方法[4]。也没有采用 Schema.org[14]或者 Facebook Graph[15]的方案，因为他们更偏向于站点数据。DataDEX 采用的是适合移动互联网和云服务的，面向应用行为的数据分类法，参考了 MSGraphAPI[5]，对移动端数据的操作和数据存取。后端数据按对象结构存储，组成用户知识图谱数据。



这些类目的数据按顺序登记，有反作弊处理机制，及时发现伪造或大批量刷数据的问题。同时建立审核委员会，对核心数据定期校验，并且对争议进行调解和仲裁。

4 DEX 方案

4.1 智能合约和 DApp

Alita Network 主链集成了自动交易和图灵完整的编程语言，由 CIYAM[10]开发人员建立的智能合约机制。 在由区块链和联合计算技术支持的网络上

以编程方式执行任务代表了数据计算的新时代。 在 DataDEX 上具有此类功能的能力可以改善整个环境，从而提高效率和可用性。

创建资金池 Dataset 时，会自动部署智能合约，存储 Dataset 和 ALITA 交易对，并且根据 AMM 算法实时价格。

我们提供在线的 UI 便捷的创建 Dataset 和 Data Token，可以选择资金池模式和纯交易模式。

4.2 做市算法

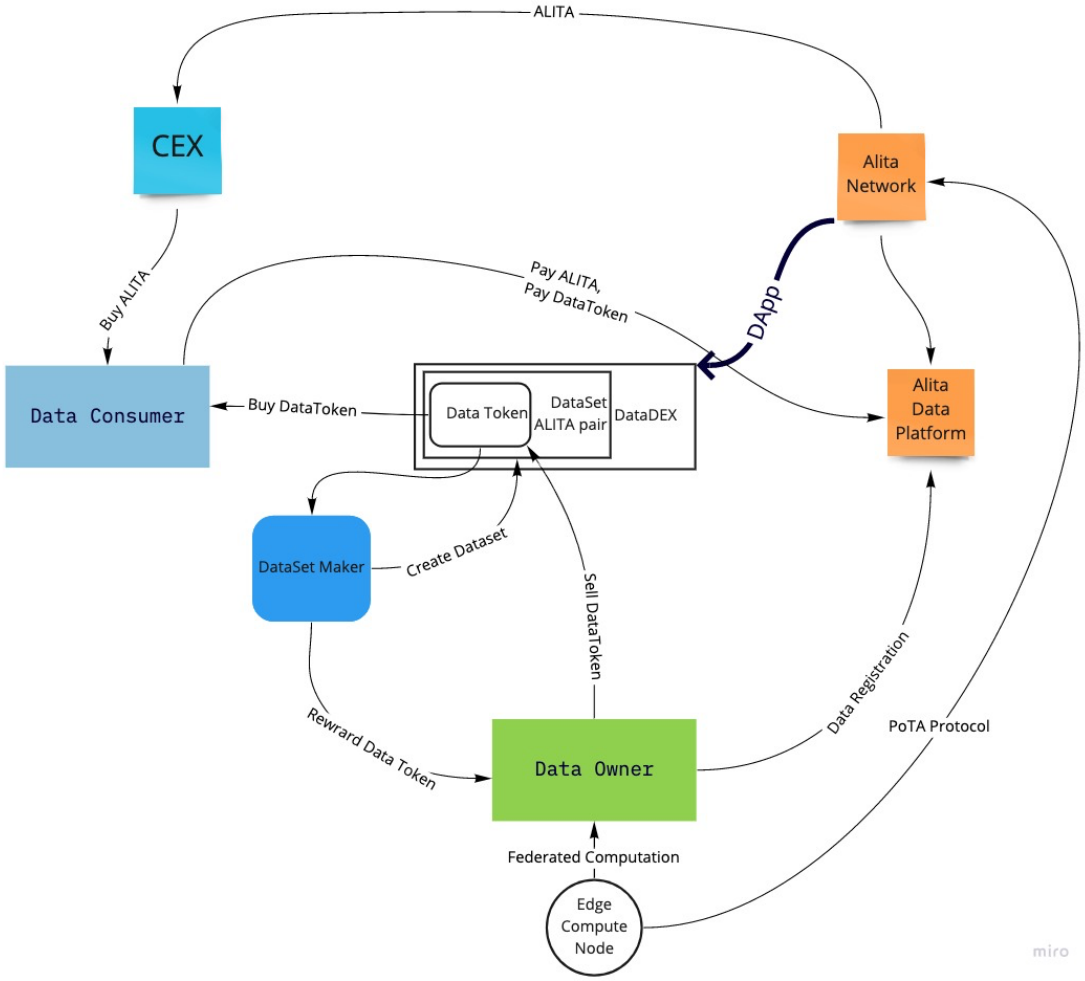
参考 Uniswap[11]的做市算法，Datatoken 和计算 token ALITA 的比乘以一个常数。

$$pt = \mathrm{K}\,\frac{\mathrm{DataToken}}{\mathrm{ALITA}}$$

我们也会为增加流动性的 Data Maker 提供流动性挖矿机制，参考了 SushiSwap[12]的 Pool 和 farming。具体收益模型见《DataDEX 经济白皮书》

数据确权时的验证需要外部数据，我们设计了预言机模式，让外部数据来佐证数据确权的真实性，Data Maker 也可以发起预言机验证活动，数据 owner 提供外部数据，超级节点作为验证者可以验证数据的合法性，通过公共数据接口，比如公安、学历认证网、芝麻信用等数据开放接口，并保证真实性。类似预言机 REST[13]的模式。由于预言机在外部数据结合方便有很好的治理能力，但无法处理大规模的数据，所以在我们的项目里用于验证数据确权合法性。

5 流程



1. 数据制作者创建新的数据令牌，并在DEX中创建数据集DataSet（Pool），然后设计数据需求，解决方案和潜在客户。在CEX中购买一定量的ALITA以建立初始流通盘，并邀请社区成员加入。
2. 社区成员以“数据提供者”的身份加入，他们将识别数据并获得数据令牌的奖励。
3. 商业数据客户，数据消费者对数据集中越来越多的数据提供者非常感兴趣，并从CEX（Bithumb，Huobi等）购买了ALITA，然后存入Dataset（池）获取数据token。随着此类用户的增加，数据token的价格将上涨。数据消费者将支付数据所有者数据令牌和ALITA作为数据费用和计算费用。
4. 数据所有者可以选择出售手头的数据令牌。根据流动性定价策略，它可能会产生数十倍的利润。
5. 数据制作者通过Data Token和ALITA之间的交易费获得收入

6 关键问题

6.1 区块链时代隐私保护的数据交换有哪些要素？

1. 支持大规模数据处理，目前全球最大规模的计算框架是MapReduce，DataDEX是基于Alita Network的大数据处理网络计算数据的
2. 支持包括手机在内的边缘端计算，充分利用联合计算的隐私保护能力，将数据留在用户主权区，不收集明文明细数据。其中谷歌的联合学习[7]与联合分析[8]就是很好的模式，但他计算模式不够通用，而且对数据资本化没有布局。
3. 链上计算，数据不能在链下计算，否则无法实现隐私保护。DataDEX利用计算资源消耗的共识PoTA[6]，和TEE[9]技术，确保共识安全，计算完整性以及可验证性。

6.2 为什么不在以太坊或者其他 PoW 公链实现这个 DEX？

如上述，隐私保护的大数据交换必须要基于一个边缘计算公链，大数据处理是共识的一部分，才能实现隐私保护的分布式计算，并且让使用权在智能合约中交易。

6.3 与其他隐私计算协议的区别

我们支持大规模数据处理，并不是智能合约内部的计算，合约内部只处理使用权交易。也不使用预言机处理数据，以保证数据处理的效率和规模。

6.4 中心化的 Data MarketPlace 为何没有成功？

一方面中心化带来隐私风险，另一方面中心化数据市场没有合适的价格发现机制，无法使供需双方对数据价值快速共识。

6.5 数据验证和确权如何防止伪造？

数据上链的都是结构化数据，不是数据包，有明确的类目体系，借鉴Microsoft Graph API 的分类规范，以个人用户为中心，抽象出个人数

据类型和行为明细，并且进行数据特征校验，确保脏数据无法上链确权。

7 结论

作为技术人员和数据科学家，我们深信个人数据潜力巨大，而且用户必须控制，从而在数据使用的风险和收益之间进行权衡。DataDEX 是提供隐私保护的个人数据资本化的一种尝试，该存储对用户而言既方便又安全拥有，管理和控制他的数据。通过去中心化数据交易所和联合计算网络为个人提供了一种重新控制其数据和隐私的新方法，同时支持创建智能的，数据驱动的应用程序。

8 参考资料

[1] 联合计算 Federated Computation，一种隐私保护的移动端边缘计算方案 https://baike.baidu.com/item/%E8%81%94%E5%90%88%E8%AE%A1%E7%AE%97/

[2] Personal Data: The Emergence of a New Asset Class, http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf.

[3] Alibaba One Data https://dt.alibaba.com/onedata.htm

[4] Dimensional Modeling Techniques: The Kimball Method, Ralph Kimball https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/

[5] Microsoft Graph API https://docs.microsoft.com/en-us/graph/overview

[6] Alita Network, Federated Computation support multiple devices including Android Phone http://alita.global/whitepaper_en

[7] Google Federated Learning https://arxiv.org/pdf/1902.01046.pdf

[8] Google Federated Analytics https://ai.googleblog.com/2020/05/federated-analytics-collaborative-data.html

[9] Open Enclave SDK, TEE SDK can across TEEs from different vendors. https://github.com/openenclave/openenclave

[10] CIYAM safety automation transaction language. https://github.com/ciyam

[11] Uniswap core V2 whitepaper https://uniswap.org/whitepaper.pdf

[12] SushiSwap Pools Docs https://help.sushidocs.com/products/sushiswap-pools

[13] NEST Protocol: A Distributed Price Oracle Network https://nestdapp.io/lib/nestwhitepaper/ennestwhitepaper.pdf

[14] Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond. http://schema.org

[15] FaceBook Graph API: https://developers.facebook.com/docs/graph-api/